

The AI Model Risk Catalog: What Developers and Researchers Miss About Real-World AI Harms

Pooja S. B. Rao^{1,3}, Sanja Šćepanović^{2,4}, Dinesh Babu Jayagopi³, Mauro Cherubini¹, Daniele Quercia^{2,5}

¹University of Lausanne, Lausanne, Switzerland

²Nokia Bell Labs, Cambridge, UK

³International Institute of Information Technology Bangalore, Bangalore, India

⁴University of Oxford, Oxford, UK

⁵Politecnico di Torino, Turin, Italy

{pooja.rao, mauro.cherubini}@unil.ch, jdinesh@iiitb.ac.in, {sanja.scepovic, daniele.quercia}@nokia-bell-labs.com

Abstract

We analyzed nearly 460,000 AI model cards from Hugging Face to examine how developers report risks. From these, we extracted around 3,000 unique risk mentions and built the *AI Model Risk Catalog*. We compared these with risks identified by researchers in the MIT Risk Repository and with real-world incidents from the AI Incident Database. Developers focused on technical issues like bias and safety, while researchers emphasized broader social impacts. Both groups paid little attention to fraud and manipulation, which are common harms arising from how people interact with AI. Our findings show the need for clearer, structured risk reporting that helps developers think about human-interaction and systemic risks early in the design process. The catalog and paper appendix are available at: <https://social-dynamics.net/ai-risks/catalog>.

Introduction

Risk and harm have been central concerns in AI safety and ethics research. Researchers have worked to identify and organize these concepts into taxonomies (Weidinger et al. 2022; Yampolskiy 2016). Some focus on Large Language Models and Generative AI (Weidinger et al. 2021, 2022; Stahl and Eke 2024), while others address broader AI systems (Yampolskiy 2016; Wirtz, Weyerer, and Sturm 2020) or Artificial General Intelligence (McLean et al. 2023). These efforts often use different definitions of risk and harm. We follow the OECD (Perset and Aranda 2024), which defines risk as the chance of harm, and harm as a risk that has caused damage. In some contexts, the terms are used interchangeably. The MIT Risk Repository (Slattery et al. 2024) compiles risks from academic work on AI frameworks and is the largest collection of researcher-identified AI risks to date. The AI Incident Database (McGregor 2021) catalogs real-world harms, which have been further classified by Velázquez et al. (2024).

Existing research has not explored how developers describe the risks of specific models or how those models might fail in typical user scenarios. This perspective is criti-

cal, as developers can offer grounded insight into the behavior and limitations of the systems they build.

To address the lack of risk data linked to specific models and described by their developers, we analyzed nearly half a million model cards (Mitchell et al. 2019) from Hugging Face as of July 2024. Model cards are a widely adopted standard for documenting AI models, adopted by both major technology companies and individual developers. This work makes two main contributions (Figure 1):

1. *Analyzing risks identified by developers to create the AI Model Risk Catalog.* We collected all available model cards on Hugging Face and using established taxonomies (Weidinger et al. 2022; Slattery et al. 2024), grouped similar risks within each category, and selected representative examples. The result is the AI Model Risk Catalog, which includes 2,863 categorized risks tied to specific models. The catalog is publicly available at: <https://social-dynamics.net/ai-risks/catalog>.
2. *Comparing developer and researcher risks with real-world harms.* We compared the types of risks emphasized in three sources: developer-described risks from model cards, researcher-identified risks from the MIT Risk Repository, and real-world harms from the AI Incident Database. Developers tend to report technical risks such as model limitations, safety issues, and bias. These account for over half the harms recorded in real-world incidents. Researchers focus on governance, societal impacts, and threats to human agency, which together explain fewer than 15% of incidents. Notably, the largest share of incidents involves malicious use and misinformation. These risks are underrepresented by both developers and researchers, likely because they depend on unpredictable human behavior: hard for researchers to anticipate, and even harder for developers to account for (Velázquez et al. 2024).

We close by discussing how our catalog complements existing efforts to map AI risk and what it means for developers, researchers, journalists, policymakers, and the public.

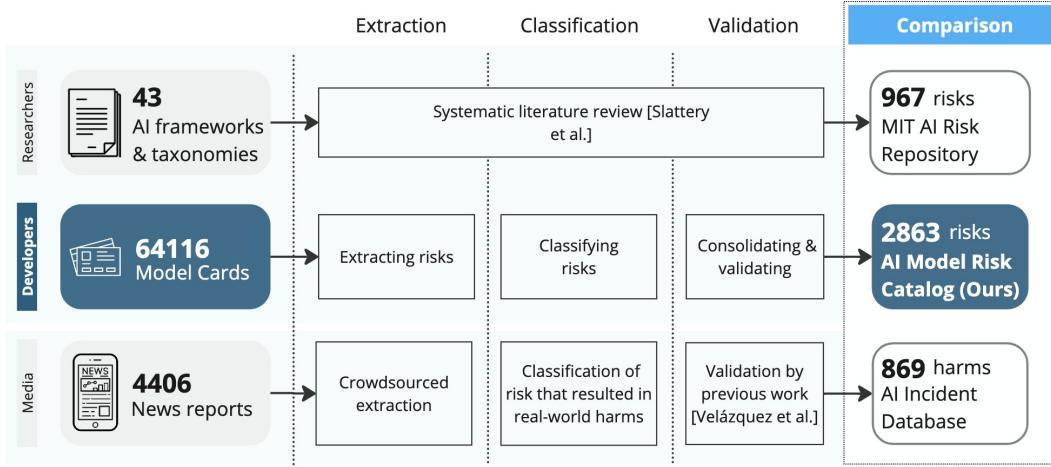


Figure 1: Overview of our methodology. Our methodology has three steps: extracting, classifying, and consolidating risks from model cards to build the AI Model Risk Catalog. We then compared the developer-reported risks in the catalog to those identified by researchers in the MIT Risk Repository (Slattery et al. 2024) and to real-world harms from the AI Incident Database (McGregor 2021). The catalog is based on 64,116 model cards with risk sections, from which we extracted 37,401 risk mentions (including duplicates). After consolidation, we identified 2,863 unique risks from 2,672 model cards, making this the largest collection of AI risks across the three sources.

Related Work

We start by reviewing research on AI risk taxonomies, followed by that on collating AI risks into databases.

AI risk taxonomies

The taxonomization of AI risks has grown significantly in recent years (Yampolskiy 2016; Liu et al. 2024b; Stahl and Eke 2024; Shelby et al. 2023). One of the most widely used taxonomies proposed by DeepMind (Weidinger et al. 2022), categorizes risks into six primary groups: *representation and toxicity*, *misinformation*, *information and safety*, *malicious use*, *human autonomy and integrity*, and *socioeconomic and environmental harms*. Each category is further divided into a total of 21 subcategories, such as *toxic content* under the first category and *environmental damage* under the last. Weidinger et al. (2023) also explore the context of harm origins, identifying three layers: *capability* (related to the system’s technical features), *human interaction* (linked to users’ experiences with the system), and *systemic impact* (including the broader context in which the system operates).

Other notable taxonomies include that by Shelby et al. (2023), which identifies five key harm categories: presentation, allocative, quality of service, interpersonal, and social system harms. Yampolskiy (2016) proposes a classification based on the causes of risks, organized along three axes: entity (human or AI), intentionality (intentional or unintentional), and timing (pre-deployment or post-deployment). Arguing that many existing taxonomies, including those mentioned above, are primarily designed for researchers and policymakers, making them less accessible to broader audience, Abercrombie et al. (2024) introduced the AI, Algorithmic, and Automation Harms Taxonomy, which aims to be more comprehensible to the general public. This framework adds categories such as physical and psychological harms.

AI risks data sources

Another important area of research focuses on collating AI risks into data sources.

Risks envisioned by researchers. In the most comprehensive effort to date, Slattery et al. (2024) analyzed 43 scholarly and industry frameworks proposing various AI risk taxonomies and compiled them into the MIT AI Risk Repository. This continuously updated repository currently lists 967 risks, categorized using two taxonomies: a causal taxonomy adapted from Yampolskiy (2016) and a domain taxonomy extended from the DeepMind taxonomy (Weidinger et al. 2021). Using a best-fit approach (Carroll et al. 2013), Slattery et al. (2024) refined the DeepMind taxonomy by adding a seventh category, *AI system safety, failures, and limitations*, and making minor adjustments to existing categories. Derczynski et al. (2023) introduced “Risk Cards” providing definitions, categorizations, and examples of harms tied to specific language models (LMs). They identified risks from scholarly literature, categorized them using the taxonomies of Weidinger et al. (2021) and Shelby et al. (2023), and released these findings as the LM risk cards starter set. This set, however, is a result of a more specific effort (i.e., focuses on LMs only), and features significantly fewer risks compared to the MIT AI Risk Repository.

Automatically envisioned risks. Tools have also been developed that often use LLMs to generate risk ideas and help AI practitioners anticipate risks during the design phase. For instance, Herdel et al. (2024) created ExploreGen, an LLM-based tool that generates potential AI uses, and categorizes their risk level according to the EU AI Act. Wang et al. (2024) developed FarSight, an interactive tool that supports AI prototyping by providing news articles on related incidents and helping practitioners explore potential risks

and affected stakeholders. Buçinca et al. (2023) introduced AHA!, a tool leveraging vignettes to describe possible harms with the help of LLMs or crowdsourcing. CardGen (Liu et al. 2024a) is a pipeline that uses retrieval-augmented generation (RAG) to complete missing sections of model cards, including risk sections, by drawing on information from papers and GitHub projects. Lastly, RiskRAG is another RAG tool assisting developers in envisioning risks of their models that combines ExploreGen, and curates data from other model cards and the AI Incident database (Rao et al. 2025).

AI risks materialized as harms. To track cases where AI systems have caused real-world harm (Buolamwini and Gebru 2018), several databases of AI incident reports have been developed (Rodrigues, Resseguier, and Santiago 2023). The most well-known is the AI Incident Database (McGregor 2021), which curates news reports on AI failures and categorizes these incidents (Turri and Dzombak 2023; Abercrombie et al. 2024).

Velázquez et al. (2024) classified incidents recorded in the incident database using the DeepMind taxonomy (Weidinger et al. 2022), finding that most incidents stemmed from the human-interaction layer, with harms most frequently falling under the human autonomy and integrity, or representation and toxicity categories. Bogucka, Šćepanović, and Quercia (2024) visualized data from the incident database, including a subset focused on mobile computing (Bogucka et al. 2024c), making the information more accessible to broader audiences. Other prominent databases include the AI, Algorithmic, and Automation Incidents and Controversies (Pownall 2023), the OECD Monitor (Organisation for Economic Co-operation and Development (OECD) 2025), and “Where in the World is AI?” (AI Global: Global AI Incident and Mapping Project 2025).

Additionally, AIES and FaccT communities have significantly contributed to studying instances of real-world harms (Ali et al. 2019; Albert and Delano 2021), and offered recommendations for auditing AI systems (Raji and Buolamwini 2019).

Research Gap

Previous work has focused on classifying AI risks into taxonomies, identifying risks envisioned by researchers, automatically suggesting risks to AI practitioners, and curating harms in the aftermath of real-world incidents. However, the valuable perspectives of AI developers has been overlooked, limiting our understanding of unique risk profiles of specific AI models and the interactions between their capabilities and common uses. It also means we lack a clear picture of how different experts (e.g., researchers and developers) think about AI risks, and how their concerns match the harms that have already occurred.

Methods

To address this research gap, we built and validated the AI Model Risk Catalog based on risks identified by developers, and compared it with risks identified by researchers in the MIT Risk Repository, and with harms reported in the AI Incident Database.

Methods for building the AI Model Risk Catalog

We first describe the model cards dataset, and then the methods of using LLMs for extracting risks from this dataset, classifying the extracted risks into taxonomies, and consolidating and validating the classified risks into a catalog.

Downloading model card snapshots from HuggingFace.

In July 2024, we obtained a snapshot of the model repository from HuggingFace using the HF Hub API¹. This included 765,973 model repositories, with 461,181 (60%) model cards. For every model card, we employed regular expressions to identify sections related to risks. Specifically, we searched for mentions of risks, limitations, bias, ethical considerations, out-of-scope uses, misuse, responsibility, and safety. This approach identified 64,116 (14%) model cards with risk-related sections. Due to the absence of standardized content requirements on HuggingFace, many model cards are incomplete, and numerous risk sections are only slightly modified replicas of one another. Among the 64,116 model cards, an overwhelming majority of risk sections (96%) were exact duplicates. We retained 2,672 model cards with unique risk-related content (selecting the card with highest download count in cases of duplicates) as our *Model Cards up to 2024* dataset. Furthermore, we also considered a snapshot of all the model cards up to October 2022, released by previous research (Liang et al. 2024) to analyse how the risks reported in model cards have evolved over the two years. This snapshot consisted of 74,970 model repositories and 32,111 (42.8%) model cards. We extracted the risk-related sections, resulting in 5,546 (17.3%) model cards with any risk-related sections. Among these, 95% were exact duplicates with 322 model cards having unique risk-related content referred to as *Model Cards up to 2022* dataset. See Table 1 for details on both data snapshots.

Extracting risk mentions. Given the scale of our dataset, manually extracting risks from 2,672 model cards was not feasible. We therefore used a large language model (GPT-4o) to automate the extraction process and validated its accuracy against a manually annotated sample. Large language models have shown strong performance in zero-shot and few-shot annotation tasks (Ziems et al. 2024; Strachan et al. 2024). We prompted GPT-4o with definitions and examples, asking it to identify whether a section discussed risks and to extract distinct mentions in a verb-object format. To reduce variability and hallucinations, we set the temperature to zero (Peeperkorn et al. 2024). We refined the prompt iteratively until its outputs matched human annotations on a separate 10% test set, achieving 90% agreement on a sample of 50 cards. All extracted risks were reviewed by the authors for consistency and accuracy. We removed exact duplicates and further grouped near-duplicates using two methods: (1) *fuzzy string matching* to detect similar phrasing, and (2) *contextual embeddings with cosine similarity*, using the *bge-large-en-v1.5* model (Muenennighoff et al. 2023). Pairs with a fuzz score of at least 75 and embedding similarity above 0.85 were treated as duplicates. We retained the longer entry

¹https://huggingface.co/docs/huggingface_hub/v0.5.1/en/package_reference/hf_api

Table 1: Statistics of two HuggingFace model card snapshots that we used for creating two versions of the AI Model Risk Catalog (2022 and 2024). Out of all the model cards, very few have a completed risk section, and an even smaller number have unique risk sections (i.e., sections that are not copied from other cards).

Dataset	Stats for all model cards			Stats for the model cards with unique risk sections	
	Total cards	With risk sections	With unique risk sections	Average # downloads	Characters in risk sections
<i>Model cards up to 2024</i>	461,181	64,116	2,672	118,508.41	759.69
<i>Model cards up to 2022</i>	32,111	5,546	322	364,639.04	882.90

in each pair. These thresholds were calibrated on 5% of the dataset.

Taxonomizing risks. After extracting risks from model cards, we classified them using two taxonomies: the Deep-Mind taxonomy (Weidinger et al. 2023) and the MIT Risk Repository taxonomy (Slattery et al. 2024). We used GPT-4o to assign each risk to the most suitable category. To guide the model, we provided definitions from both taxonomies and asked it to match each risk accordingly. We refined the prompt until its outputs matched manual annotations on an unseen 10% sample. Because the model’s outputs vary, we classified each risk three times and kept only the categories that appeared at least twice. To evaluate accuracy, we compared the model’s results against a manually annotated sample of 50 risks and found 84% agreement.

Consolidating risks. A manual validation of a sample of extracted risks revealed that the automatic duplicate removal approach was partially effective. To enhance the dataset’s credibility, the extracted risks were manually reviewed and refined. For each sub-category of the MIT risk taxonomy, the first author assessed the risks within the category, identified those with significant similarity, and removed duplicates, prioritizing the retention of the risk with the most comprehensive information. These eliminations were then independently verified by the second author. Any disagreements were resolved through discussion. Care was taken to preserve granular details, such as risks specific to particular models or datasets, as well as risks that represented instances of other risks but included additional information. This process resulted in our AI Model Risk Catalog.

Validating classification of risks. To validate the quality of the risk classifications generated by the LLM, we evaluated its performance on the manually coded MIT risk repository. We applied our custom prompt to extract the LLM’s classifications of risk mentions in the repository. These predictions were then compared against the repository’s manual coding, which served as the ground truth. The results showed an accuracy of 83% and a macro-averaged F1 score of 81% for the seven-class classification task. Most misclassifications occurred in the predicted categories of *malicious actors and misuse* (32 risks), *misinformation* (26 risks), and *AI system safety, failures, and limitations classes* (22 risks). Upon closer inspection, we found that these were not necessarily errors. As noted by Weidinger et al. (2023) and Slattery et al. (2024), these categories are not mutually exclusive, and many risks span multiple categories due to their

interconnected nature. For example, risks “the demonstrated ability of anonymous actors to accumulate resources online (e.g., Satoshi Nakamoto as an anonymous crypto billionaire)” and “this is the risk posed by an ideal system if used for a purpose unintended by its creators” are classified by LLM under *malicious actors and misuse*, but are found under *AI system safety, failures, and limitations* in the repository. Obviously, the unintended, anonymous, and negative use mentions also justify their inclusion under *malicious uses* category. Similarly, “AI-generated or synthesized content can lead to the spread of false information, discrimination and bias, privacy leakage” could be both risk of *misinformation* (as classified by the LLM) and *discrimination and toxicity* (MIT groundtruth).

To ensure that the high classification accuracy translates to our data, we conducted a manual review of how all risks are classified in our catalog, paying special attention to the overlapping categories discussed above. The review revealed less than 1% misclassifications, validating the accuracy and quality of our catalog.

Thematically analyzing risks. We conducted a thematic analysis (Braun and Clarke 2012, 2006) of the extracted risks using inductive coding where one author coded the data to comprehend and highlight the characteristics and quality of risk reporting. These codes were then jointly discussed by two authors and resolved for any disagreements.

Methods for comparing risk sources

We downloaded two prominent sources of AI risks, and implemented methods for comparing our catalog to them.

Sources of researcher-identified risks and real-world harms. We used two public sources to compare developer-reported risks with those identified by researchers and with real-world harms. First, the MIT Risk Repository² (Slattery et al. 2024) compiles a structured repository of 967 risks drawn from 43 AI risk frameworks, categorized across two taxonomies. Second, the AI Incident Database (AIID)³ catalogs cases where AI systems have caused harm or failed in practice. Reports are submitted by contributors, reviewed by volunteer editors, and grouped by incident across multiple media sources. As of January 2025, it included 869 incidents based on 4,406 media reports. For brevity, we refer to these sources as the repository (MIT Risk Repository) and the database (AIID) throughout.

²<https://airisk.mit.edu/>

³<https://incidentdatabase.ai>

Classifying risks and harms into taxonomies using LLM.

For AI Incidents, we followed the method of Velázquez et al. (2024), classifying each incident description into our two chosen taxonomies. As with model card risks, we provided definitions of the layers and categories from both taxonomies in the prompt. Each incident description was classified three times, and categories that appeared at least twice were kept to ensure consistency. The exact prompts are in Appendix *Prompts* in the pre-print version of this paper. For the MIT Risk Repository, we used the manual labels already provided for the MIT taxonomy and applied the same prompting procedure to map each risk to the DeepMind taxonomy. This process gave us common coding across all three sources (model cards, MIT repository, and AIID), letting us compare how risks envisioned by developers and researchers aligned or diverged from those shown as harms in the news.

Comparing risk sources. To compare the prevalence of risk categories across the three sources, we calculated confidence intervals for the difference of proportions between each category using the Miettinen-Nurminen asymptotic score method (Fagerland, Lydersen, and Laake 2015, p. 250). This analysis employed the `diffscoreci()` function from the `PropCIs` package in R (Scherer 2018). Following prior research (Salehzadeh Niksirat et al. 2023), we chose this method over Z-tests, as the boundary probabilities (close to 0 or 1) in some categories render Z-tests and their confidence intervals unreliable (Agresti 2011, p. 164).

AI Model Risk Catalog

We built the catalog using two snapshots of Hugging Face model cards dated through 2022 and 2024. We then examined how many risks developers report, how they report them, what kinds of risks they describe, and how their reporting has changed over time.

How many risks developers report? In the snapshot of model cards up to 2024, of the 64,116 cards with risk-related sections, 54,448 (approximately 85%) lacked substantive risk information. Most of these either simply retained the default template requesting risk details (52,983), or referred to another related card for risk content. The remaining 9,668 cards contained a total of 37,401 standardized risk mentions, including duplicates.

We removed duplicates by keeping the most downloaded model cards, resulting in a final dataset of 2,672 cards with unique risk content. From these, our extraction method produced 3,588 distinct, standardized risk mentions. The consolidation of risks (including de-duplication and streamlining) reduced the preliminary set of 3,588 risks to the final set of 2,863 risks (~ 20% reduction).

In the snapshot of model cards up to 2022, of the 4,546 cards with risk-related sections, 367 lacked risk information. The remaining 4,179 cards contained a total of 9,645 standardized risk mentions, including duplicates. Removing duplicates resulted in 322 cards with 474 standardized unique risk mentions.

How developers report risks? The thematic analysis of all the extracted risks revealed five themes related to risk

reporting practices of developers:

1. *Repetition and redundancy*: Although expressed in many different phrasings, the risk mentions tend to cover similar ground, indicating pervasive concerns among AI developers. Those include bias and fairness, output quality and accuracy, safety and harmful content, privacy and security, operational and technical limitations, as well as some ethical and societal implications.
2. *Ambiguity*: In line with previous research (Bhat et al. 2023; Crisan et al. 2022a), many risk mentions are ambiguous, lack specificity, or generally addresses a large type of models. Some examples are “is not immune from issues that plague modern large language models”, “generates confusion”, “makes mistakes”, and “increases risk to users if used irresponsibly”.
3. *Granularity levels and specificity*: In addition to the risk mentions such as those above that are very general, others are very capability-specific. For instance, for a general risk saying that the “model underperforms on out-of-distribution data,” specific versions could be saying that it does so with particular languages, dialects, input types (e.g., “jpeg artifacts”), or for certain classification labels (e.g., emotion “fear”).
4. *Interdisciplinary concerns*: The risks span across multiple dimensions of AI use, from code generation and image synthesis to language understanding and ethical decision-making, highlighting that developers understand that these challenges are not isolated but rather systemic and interconnected.
5. *Warnings for deployment*: There is an overall emphasis on caution, with many entries explicitly advising against using the models in critical or unvetted applications without robust human oversight, rigorous testing, and additional safety mechanisms. Examples include “should not be used as a substitute for professional legal advice”, “should not be solely relied upon for real-time critical medical decisions.”

What types of risks developers report? To understand the types of model risks developers imagine, and align with prior work (Slattery et al. 2024; Weidinger et al. 2021), we opted to classify the risks using established taxonomies. The DeepMind taxonomy (Weidinger et al. 2023)—currently the most widely cited in the literature, as noted by Slattery et al. (2024)—served as a natural starting point. Additionally, we found that an augmented version of this taxonomy, developed as part of the MIT Risk Repository efforts (Slattery et al. 2024), introduced an extra category (*AI system safety, failures, and limitations*) applicable to many of the risks in our catalog. Moreover, this augmented taxonomy facilitates a direct comparison between our findings and those in the MIT Risk Repository, currently the largest AI risk repository. For these reasons, we chose the MIT risk taxonomy as the second one to classify our risks with.

Appendix Figure 5 show how the cataloged risks map onto the MIT risk taxonomy. The most common category is *discrimination and toxicity* (44%), with 24% of all risks tied to *unfair discrimination and misrepresentation*, fol-

lowed by *exposure to toxic content* (8%) and *unequal performance across groups* (8%). These risks often refer to biased training data and skewed model outputs on fairness benchmarks (Le Quy et al. 2022). The next most frequent category is *AI system safety, failures, and limitations* (37%), with most of these risks falling under *lack of capability or robustness* (35%). They include reports of model bugs, inefficiencies, or errors in output quality, reflecting a primarily technical view among developers. The *misinformation* category focuses on hallucinated or inaccurate information. *Privacy and security* risks relate to code vulnerabilities, unauthorized sharing of copyrighted material, and memorization of training data. Though less common, *malicious actors and misuse* includes concerns about models enabling illegal activity or violating human rights. The *overreliance and human agency issues* category addresses AI use in sensitive domains like healthcare and the risks of anthropomorphizing models, though it accounts for less than 2% of all risks. Similarly, the *socioeconomic and environmental harms* category—also under 2%—covers the use of computational resources, legal compliance, and broader societal impacts.

These risk patterns reflect the dominance of language models on Hugging Face (62%), which are widely known for raising concerns about discrimination and misinformation (Bender et al. 2021; Ousidhoum et al. 2021). However, our analysis by input and output modality, shown in Appendix Table 2, reveals more nuances and details. Non-text inputs are more often linked to *socioeconomic and environmental harms* (22% of such risks), and *privacy and security* (20%). In total, 9% of all risks involve non-text inputs, and 17% involve non-text outputs. These risks span thousands of models and reflect a known gap in safety evaluation for non-text modalities (Weidinger et al. 2023). The risks identified in our catalog could serve as an inspiration or starting point while addressing this gap, and could help guide future work in this area.

We also find that different risk categories become more prominent with *mimodal* models. For models with multimodal input, compared to other risk categories, developers report more *malicious use* (14% of that category), as well as *privacy and security* risks (13%). These findings show how risk profiles shift when looking beyond text-based models.

As shown in Appendix Figure 6, the distribution of risks across DeepMind’s taxonomy reflects similar main trends described when using the MIT taxonomy. However, because the DeepMind taxonomy lacks a category for model-specific risks—*AI system safety, failures, and limitations*—some risks in our catalog could not be clearly classified and were instead all grouped under *representation and toxicity harms*, which in this case accounts for over 62%. For more clarity, and easier comparison with the MIT Risk Repository, we report the rest of our findings in main text using the MIT taxonomy.

How developers’ risk reporting has evolved? Previous research analyzing the 2022 snapshot of model cards (Liang et al. 2024) reported that developers often struggle to complete the risk sections: less than 17% of cards had any risk content reported. To understand if risk reporting has changed

since then, in Figure 2, we compare the 2022 and 2024 model card snapshots, and our two versions of the AI Model Risk Catalog based on those snapshots. While there is a significant increase in the total number of model cards over this period, the percentage of cards with completed risk sections has *decreased* to 14%, as has the percentage with unique risk sections (< 1%). We speculate this decrease happened because newer models are often derived from or fine-tuned versions of foundational models like GPT, and may end up reusing their original risk sections. Additionally, the average length of the risk sections has also declined (see Table 1). Our findings not only confirm those of earlier studies that highlight developers’ struggles with risk communication (Liang et al. 2024; Bhat et al. 2023; Crisan et al. 2022b), but they raise a critical concern that the situation may have even worsened over time.

In terms of the types of risks developers envision, Results in Figure 2 indicate a relative stability overall. The notable difference is a swap in the two most frequent categories: in 2022, *AI system safety, failures and limitations* were most prevalent, followed by *discrimination and toxicity*; in 2024, this order has reversed. Also, the number of *misinformation* risks has significantly reduced in 2024 (threefold decrease), whereas those about *malicious actors and misuse* have increased (fourfold increase, though still accounting for only 4% of the total risks), as well as those about *privacy and security* (sixfold increase, accounting to 3% of the total risks now). This is likely related to the increase in the multimodal models (see (Yin et al. 2024), and our Appendix Figure 4), which are linked with more of such risks, as we discussed in the previous section. The prevalence of risks in other categories has not changed significantly.

Comparison of risks envisioned by AI developers and researchers with harms recorded in the news

Having built the AI Model Risk Catalog, we compared risks reported by developers, those identified by researchers, and harms recorded in real-world incidents. Figure 3 shows the results, we highlight three key takeaways next. Note that, to capture the full range and frequency of developer-reported risks, we analyzed all 37,401 mentions, including duplicates, from 64,116 model cards with risk sections, rather than limiting our analysis to the 2,863 unique entries in the catalog. This approach aligns with the MIT Risk Repository and the AI Incident Database, where the same risks or harms may appear in multiple research papers or news reports.

Takeaway 1: Developers imagine more of the risks tied to AI capabilities (37%) and bias (over 44%) than their share in real-world recorded harms (24% and 27%). Still, these risks account for over half of the harms, which supports the relevance of developers’ concerns.

When developers focus on risks tied to model capabilities (*AI system safety, failures and limitations*, 37%), it reflects their hands-on role: they build, test, and evaluate model performance. They tend to report concrete technical issues (e.g., “misclassification due to tokenization,” “context window

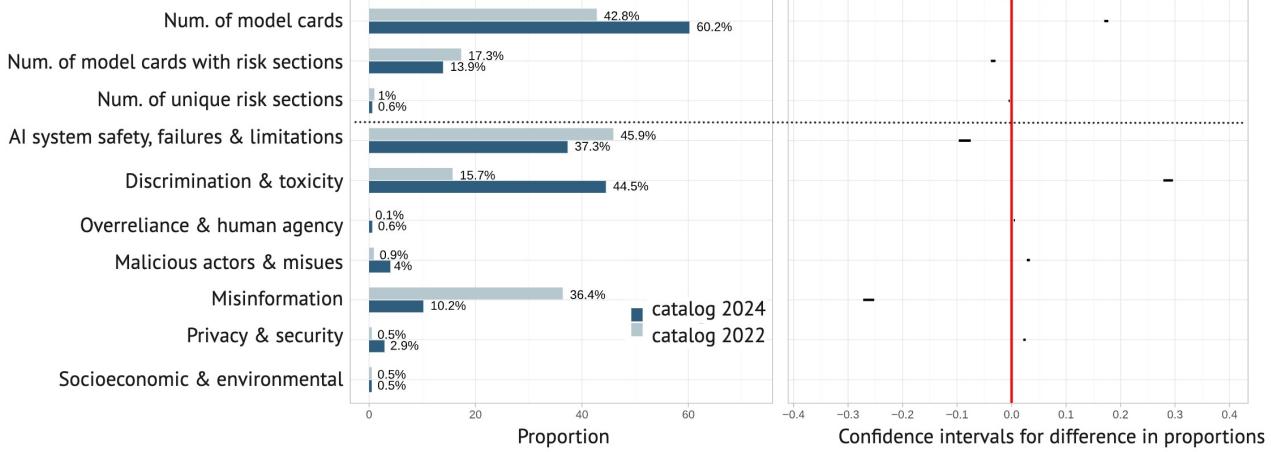


Figure 2: Comparing AI Model Risk Catalogs from 2022 and 2024. *Left:* The percentages of risks in the catalog that belong to each category. *Right:* 95% confidence intervals for the pairwise differences between each pairs of percentages. A greater distance of the interval from the vertical red line indicates a larger difference, while a narrower interval reflects higher certainty. The top section (above the dashed line) shows statistics from the model cards snapshot, while the bottom section presents the seven risk categories. Although the number of models with model cards has grown significantly, the proportion with risk sections has dropped even further (from 17% to 14%). The *AI system safety* risks were the most prevalent in 2022, overtaken by *discrimination and toxicity* in 2024. There is a threefold decrease in the proportion of *misinformation*, and fourfold increase in the *malicious uses* risks between the two years. The remaining categories have remained mostly stable.

saturation,” or “texts longer than 128 tokens”) along with problems in how models process information, the quality of outputs, and behavior in specific settings. Unlike risks in the research repository or the incident database, the developer-reported catalog highlights fine-grained model behaviors in controlled environments. As shown in Appendix Figure 5, most of these developer-imagined risks fall under the sub-category of *lack of capability or robustness*, which also accounts for most real-world harms. The other two sub-categories—*lack of transparency or interpretability* and *AI pursuing its own goals*—each account for less than 1% of developer-reported risks and real-world harms. Still, this points to a potential gap: transparency and interpretability are key concerns in responsible AI (Barredo Arrieta et al. 2020; Samek, Wiegand, and Müller 2017), especially when it comes to the adoption in high-stakes domains such as healthcare (Amann et al. 2020). Notably, researchers give these risks slightly more attention (2.6%), as shown in the MIT Risk Repository distribution in Appendix Figure 3.

Bias-related risks (*discrimination and toxicity*) make up the majority (over 44%) of developer-reported risks and 27% of real-world harms. This shows that developers are not only concerned with model capability, but also with value alignment, especially around fairness and the risk of producing discriminatory or toxic content. As discussed earlier (see Figure 2), these concerns have more than doubled since 2022, showing that developers are paying increasing attention to discrimination from their models. These risks are often tied to specific model types and modalities, such as speech-to-text or audio-visual models. Developers also anticipate use-specific risks: bias in models used in health-

care, hiring, or other sensitive areas is commonly mentioned. Some catalog entries are highly specific; for example, inherited bias from a certain dataset or “biases toward anime female characters.” Among the three bias-related sub-categories (Appendix Figure 5), most developer-reported risks fall under *unfair discrimination and misrepresentation* (24%), followed by *exposure to toxic content* (12%) and *unequal performance across groups* (8%).

Since 24% of recorded harms relate to *AI system safety* and 27% to *discrimination and toxicity*, the concerns raised by developers are valid and timely. One example of the first type of harm involves a warehouse robot that ruptured a can of bear spray and injured workers (incident 2). An example of the second type is when Google Image search showed a Barbie doll as the first female “CEO”—after 11 rows of male CEOs (incident 18).

Takeaway 2: Compared to recorded real-world harms, AI researchers imagine more of the risks from three categories: *socioeconomic and environmental harms* (18%), *overreliance and human agency* (7%), and *privacy and security* (12%). These risks appear nearly three times more often in research than in incident data, and over ten times more than in developer-envisioned risks.

Only 4% of real-world harms fall under *socioeconomic and environmental harms*, compared to 18% of researcher-imagined risks. We see this strong focus by researchers not as misplaced but as forward-looking. As noted by Velázquez et al. (2024), systemic harms often take longer to materialize and are less likely to be reported in the news—not because they are less serious, but because their effects are slower and

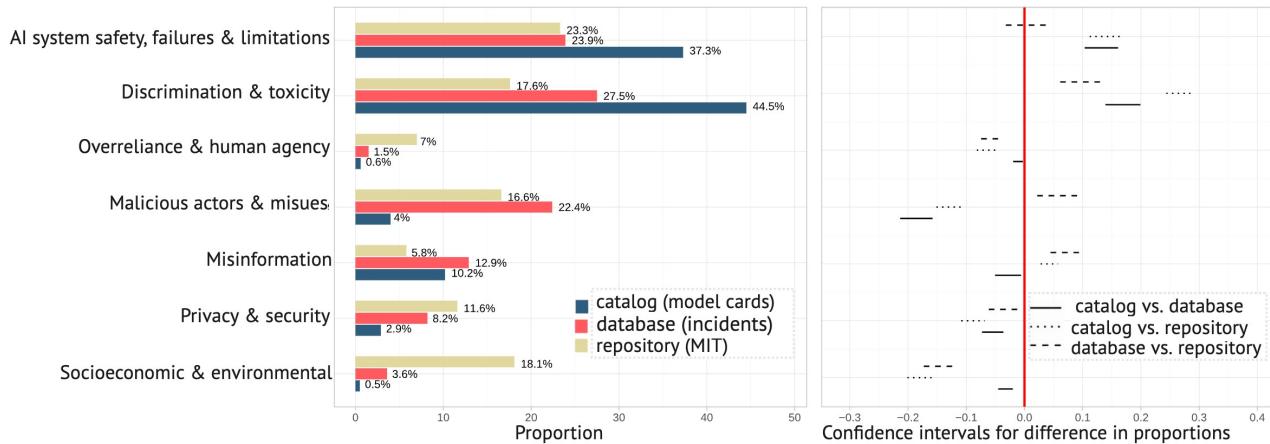


Figure 3: Comparison of three risk sources. We compare our catalog (developer-envisioned risks in model cards), repository (researcher-envisioned risks in the MIT Repository), and the database (harms recorded in the AI Incidents). *Left:* The percentages of risks in the catalog / database / repository that belong to each category. *Right:* 95% confidence intervals for the differences between each pair of percentages. A greater distance of the interval from the vertical red line indicates a larger difference, while a narrower interval reflects higher certainty.

harder to trace. Researchers give roughly equal weight (Appendix Figure 5) to subcategories such as *power centralization and unfair distribution of benefits* (4%), *increased inequality and decline in employment quality* (3%), *economic and cultural devaluation of human effort* (2%), and *environmental harm* (3%). By contrast, developers imagine all these risks rarely—between 0.1% and 0.3%.

For *overreliance and human agency*, researchers envision 7% of their reported risks in this category, while it makes up only 2% of real-world harms, and 0.6% of developer-imagined risks. Researchers focus on both *overreliance and unsafe use* (4%) and *loss of human agency and autonomy* (3%). These risks may signal early signs of systemic issues—for example, overreliance leading to new types of mistakes at work, or loss of agency affecting worker well-being. Hence, once more, we see researchers’ attention to these concerns as justified and forward-thinking.

The higher share of *privacy and security* risks in research is also notable (12% compared to 8% of real-world harms and only 3% of developer-imagined risks). While researchers envision equal proportions (6%) of *AI system security vulnerabilities and attacks* and *compromise of privacy by leaking sensitive information*, it is the latter category that results in more incidents than the former (7% vs. 1%).

Takeaway 3: Both researchers and developers tend to overlook risks tied to malicious use and misinformation. Harms of this kind linked to human interaction and social engineering are significantly more common in incidents than in expert reports, indicating blind spots among the experts.

Malicious uses of AI are widespread, accounting for over 22% of recorded harms in the incident database, yet they receive less attention from researchers (17%) and especially developers (4%). The most common subcategories are *fraud, scams, and targeted manipulation* (15%) and *dis-*

information, surveillance, and large-scale influence (7%). One example of fraud and manipulation is the use of deepfake videos to scam Canadian immigrants out of thousands of dollars (incident 876). A case of political disinformation involves a deepfake that falsely showed U.S. Congressman Rob Wittman endorsing military support for Taiwan’s Democratic Progressive Party ahead of the 2024 election (incident 876). These harms are often tied to specific model features (Charfeddine et al. 2024; Golda et al. 2024) and their data sources (Liu et al. 2025). Hence, the low number of developer-identified risks in this area shows a clear blind spot. One example from the catalog is the risk “outputs realistic faces,” linked to a text-to-video model. This risk could cause the above mentioned harms. But this description is broad and vague—it could apply to many harms and is not helpful for mitigation. A more detailed researcher-identified risk notes that “GenAI can produce images of people that look very real, as if they could be seen on platforms like Facebook, Twitter, or Tinder. Although these individuals do not exist in reality, these synthetic identities are already being used in malicious activities.” However, as with the developer risk, there are many possible misuses of GenAI that this risk could enable, which likely explains the gap between the proportions of actual harms and expert-envisioned risks in this category.

Although there is a growing body of research on *misinformation* (Chen and Shu 2024; Liu, Sheng, and Hu 2024; Garimella and Eckles 2017), the actual harms seen in the world (13%) suggest that these risks are more serious and more frequent than the developers (10%), and especially researchers (6%) envision. Harms of both of these types can be linked to human interaction (Velázquez et al. 2024), and social engineering (Wang et al. 2021; Schmitt and Flechais 2024) highlighting at a gap in expert focus.

Discussion

By curating risks from nearly half a million model cards into an *AI Model Risk Catalog*, we help clarify how developers see AI risks. We also compare the focus of developers and researchers with real-world harms reported in the news, revealing where they overlap and differ, and where important gaps remain. Below, we outline the implications of our work for bridging the divide between these groups and other stakeholders, such as journalists, policy-makers, auditors, and the wider public.

Theoretical implications

Definition and format of AI risk. Our comparison of risks from two leading sources (the repository and the incident database) with those in our new catalog shows there is no standard way to report AI risks. Some risks are described in long, narrative statements that cover not just failures or incidents but also broader consequences—such as ethical dilemmas, moral challenges, or legal issues (mainly in the repository). Others are brief and sometimes vague statements about what could go wrong (often in the catalog). Real-world harms, by contrast, are tied to specific uses, domains, and subjects harmed (as in the database).

This points to a clear need for the responsible AI community to adopt a structured standard for defining and communicating risks. Drawing on insights from all three sources, we argue that a robust risk format should at minimum specify the situation and context in which harm might occur. Without this, a single expert-defined risk can lead to many different harms of varying severity and likelihood (as we have discussed in *Results*). Such a standard would enable interoperability across sources and help all stakeholders—developers, policy-makers, auditors, and users—understand and act on risk information. A recent step in this direction is the Risk Card format (Derczynski et al. 2023) for language model risks. Still, further simplification and user studies are needed to ensure risk communication meets the needs of different audiences.

Unified source of AI risks and harms. Our results show that these three resources complement each other: the repository and database provide depth of insight and examples of broader, systemic harms, while the catalog together with risk links to specific models and datasets, can serve as a granular reference tool for the systematic auditing of AI system capabilities (Uuk et al. 2024). Together, these resources create a shared knowledge base that can support collaboration across a wide range of stakeholders.

However, while we cover two key expert groups and the media, important perspectives are still missing. These include the views of end users and affected communities, civil society organizations and NGOs working in digital rights, consumer protection, and human rights, as well as legal experts, ethicists, and sectoral specialists from fields like healthcare, finance, and energy. Involving these groups is crucial for building a fuller, more inclusive understanding of AI risks and ensuring that risk reporting and mitigation efforts meet the needs of all those affected.

Practical implications

Researchers and scholars. Our findings show that researchers have driven the discussion on the broader societal, environmental, and governance impacts of AI. However, they should place greater focus on risks linked to malicious use, misinformation, discrimination, i.e., especially those rooted in human interaction and social engineering (Schmitt and Flechais 2024). While there is a strong body of work on misinformation (Chen and Shu 2024; Liu, Sheng, and Hu 2024) and bias (O’Connor and Liu 2024; Whittaker et al. 2019; Dai et al. 2024; Yu et al. 2024), these issues remain urgent due to the frequency and severity of real-world harms, as shown by incident data. Social engineering, in particular, exploits human psychology to deceive individuals and groups, often leading to significant harm (Wang et al. 2021).

To address these challenges, researchers should develop new frameworks and practical solutions for AI risks rooted in human behavior and social dynamics. This requires closer collaboration with the human-computer interaction (HCI) community (Tahaei et al. 2023), and experts in social and behavioral sciences (Washo 2021). Such partnerships can help anticipate new threats of these types, and design interventions grounded in real user experience.

Developers. We found that model developers, much like researchers, often overlook risks related to human interaction, social engineering, malicious use, and misinformation. However, developers also show a distinct blind spot for privacy and security risks. This lack of attention is especially puzzling since privacy, security, and malicious use are directly tied to specific model designs, underlying data, and the amplified vulnerabilities introduced by LLMs (Charfedine et al. 2024; Liu et al. 2025; Bullwinkel et al. 2025).

Bridging these gaps will require developers to systematically consider a broader set of risks, both technical and human-centered. First, using risk taxonomies as assessment guides or “cheatsheets” can support more complete and systematic developer risk assessments. Second, we recommend drawing lessons from established cybersecurity and privacy practices and approaching AI systems as software that can be compromised. Security information sharing and the adoption of security-first and privacy-first approaches—as advocated by the NIST AI management framework (NIST 2024, 2023)—are key strategies (Uuk et al. 2024). Third, developers can also benefit from tools like RiskRAG (Rao et al. 2025), which guide risk thinking in relation to real-world AI uses. Future automatic tools for envisioning risks should ground suggestions in evidence from our catalog, repositories like the MIT AI Risk Repository, incident databases, and structured resources such as BenchmarkCards (Sokol et al. 2024), rather than relying solely on LLMs.

Media professionals. Our findings can help media professionals reflect on how they cover AI risks and harms. Rather than focusing only on high-profile or sensational incidents, media outlets can broaden their reporting to include a wider variety of risks and affected groups. By highlighting both common and overlooked harms, media professionals can play a key role in informing researchers and the public, shap-

ing balanced discourse (e.g., by following UNESCO’s guide to best practices in AI reporting (Jaakkola 2023)), and holding developers and policymakers accountable.

Policy makers. Policymakers can benefit from these risk sources by drawing on risks identified by each of developers, researchers, and in real-world incidents when shaping legal frameworks (Golpayegani et al. 2024). Consolidating diverse risks in this way helps pinpoint areas needing urgent attention and can guide resource allocation. These structured resources support the development of targeted regulations, compliance tools (Bogucka et al. 2024a; Golpayegani et al. 2024), and effective mitigation strategies (Uuk et al. 2024).

Unlike less detailed sources, our catalog links risks to specific models, making it more useful for compliance checks and integration into compliance platforms. Our catalog also tracks how often each risk is reported, highlighting the most common concerns across different model types and modalities. Notably, as multimodal models become more widespread, they introduce new safety challenges (Weidinger et al. 2023; Hameleers et al. 2020), and our results show these models are rapidly increasing in number among the open-source ones on Huggingface platform—an area that requires special attention from policymakers. Even though many HuggingFace models lack risk documentation, sharing model-specific risks in a public repository like ours can aid model selection, mandatory impact assessments (Bogucka et al. 2024b,a), and inform product choices for end users (Bogucka, Šćepanović, and Quercia 2024). Lastly, by adopting and encouraging open risk reporting standards, policymakers can further promote transparency and public trust in AI.

Public. Human-interaction and social engineering risks deserve special attention, and preventing these harms requires greater public awareness. The risk sources such as our catalog can help the public understand how AI might affect them (Bogucka, Šćepanović, and Quercia 2024), support informed choices about adopting technology, and encourage accountability among developers, organizations, and policymakers. Public engagement in risk reporting, e.g., through participatory auditing, and discussion can strengthen accountability, and improve AI integration into society.

Limitations

Limitations and biases of the risk sources. We treated the MIT Risk Repository as a source of risks identified by researchers, the AI Model Risk Catalog as reflecting risks described by developers, and the AI Incident Database as evidence of real-world harms. These categories are not clear-cut: many developers are also researchers, and their roles often overlap. The repository includes only a subset of academic work; most model cards do not mention risks, limiting developers’ perspectives; and the incident database may reflect media coverage biases (Shaikh and Moran 2024; Brennen 2018). While Hugging Face is one of the most comprehensive public sources of model documentation, our catalog includes only its model cards. It excludes risks described in

other platforms, such as GitHub repositories or internal company records (Fang et al. 2020). Finally, the catalog reflects a snapshot from July 2024. It does not capture model cards added or revised since then. Our findings should be read in light of ongoing updates to Hugging Face.

LLM shortcomings. Generated risks may involve inaccuracies from LLM hallucinations (Mittelstadt, Wachter, and Russell 2023). To mitigate this, we extracted and minimally standardized risk sentences to limit misrepresentation. Potential biases in LLM-based classifications (Lucioni et al. 2024) were addressed by repeating classifications thrice and applying majority agreement, with manual validation ensuring reliability.

Adverse Impact Statement

The proposed risk catalog has the potential for dual-use. Any documentation reporting failure modes and risks could be leveraged by malicious actors to scale harmful or dangerous applications of AI systems. However, the risks included in the catalog are derived from publicly available sources, i.e., model cards, and do not introduce new or undisclosed vulnerabilities. By consolidating and categorizing these risks, our aim is to support responsible AI development and risk mitigation, while minimizing the potential for misuse through careful curation and transparency.

Ethical Considerations Statement

This work adheres to ethical research principles and community norms, ensuring compliance with applicable laws and professional ethical codes. The data used in our study were sourced entirely from publicly available model cards on HuggingFace and established repositories like the MIT Risk Repository and AI Incident Database. HuggingFace allows the use, display, publication, reproduction, distribution, and creation of derivative works according to their terms of service. We did not collect or utilize sensitive user data, nor did we interact with users or deploy systems during the research.

To mitigate ethical concerns, we took several precautions. First, we acknowledge the potential dual-use risks of documenting AI failure modes and risks, which could be exploited for harmful purposes. However, all risks in our catalog were derived from publicly accessible data, and no new or undisclosed vulnerabilities were introduced. Second, we addressed biases and inaccuracies inherent to LLM outputs through repeated classification runs, majority voting, and manual validation. Finally, we made efforts to ensure the curated catalog contributes to responsible AI development by consolidating risks in a way that supports transparency and accountability.

References

- Abercrombie, G.; Benbouzid, D.; Giudici, P.; Golpayegani, D.; Hernandez, J.; Noro, P.; Pandit, H.; Paraschou, E.; Pownall, C.; Prajapati, J.; et al. 2024. A collaborative, Human-Centred taxonomy of AI, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*.
- Agresti, A. 2011. Score and Pseudo-Score Confidence Intervals for Categorical Data Analysis. *Statistics in Biopharmaceutical Research*, 3(2): 163–172.
- AI Global: Global AI Incident and Mapping Project. 2025. Where in the World is AI? Online at <https://map.ai-global.org/> Accessed: August 2025.
- Albert, K.; and Delano, M. 2021. This whole thing smacks of gender: algorithmic exclusion in bioimpedance-based body composition analysis. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 342–352.
- Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How Facebook’s Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–30.
- Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V. I.; and Consortium, P. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20: 1–9.
- Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Bhat, A.; Coursey, A.; Hu, G.; Li, S.; Nahar, N.; Zhou, S.; Kästner, C.; and Guo, J. L. 2023. Aspirations and Practice of ML Model Documentation: Moving the Needle with Nudging and Traceability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. Hamburg Germany: ACM. ISBN 978-1-4503-9421-5.
- Bogucka, E.; Constantinides, M.; Šćepanović, S.; and Quercia, D. 2024a. AI Design: A Responsible AI Framework for Impact Assessment Reports. *IEEE Internet Computing*.
- Bogucka, E.; Constantinides, M.; Šćepanović, S.; and Quercia, D. 2024b. Co-designing an AI impact assessment report template with AI practitioners and AI compliance experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 168–180.
- Bogucka, E.; Šćepanović, S.; and Quercia, D. 2024. Atlas of AI Risks: Enhancing Public Understanding of AI Risks. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 12, 33–43.
- Bogucka, E. P.; Constantinides, M.; De Miguel Velazquez, J.; Scepanovic, S.; Quercia, D.; and Gvirtz, A. 2024c. The Atlas of AI Incidents in Mobile Computing: Visualizing the Risks and Benefits of AI Gone Mobile. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction*, 1–6.
- Braun, V.; and Clarke, V. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2): 77–101.
- Braun, V.; and Clarke, V. 2012. Thematic Analysis. In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, APA Handbooks in Psychology®, 57–71. Washington, DC, US: American Psychological Association. ISBN 978-1-4338-1005-3.
- Brennen, J. 2018. An industry-led debate: How UK media cover artificial intelligence. *Reuters Institute for the Study of Journalism*.
- Buçinca, Z.; Pham, C. M.; Jakesch, M.; Ribeiro, M. T.; Olteanu, A.; and Amershi, S. 2023. Aha!: Facilitating ai impact assessment by generating examples of harms. *arXiv preprint arXiv:2306.03280*.
- Bullwinkel, B.; Minnich, A.; Chawla, S.; Lopez, G.; Pouliot, M.; Maxwell, W.; de Gruyer, J.; Pratt, K.; Qi, S.; Chikanov, N.; et al. 2025. Lessons From Red Teaming 100 Generative AI Products. *arXiv preprint arXiv:2501.07238*.
- Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.
- Carroll, C.; Booth, A.; Leaviss, J.; and Rick, J. 2013. “Best fit” framework synthesis: refining the method. *BMC medical research methodology*, 13: 1–16.
- Charfeddine, M.; Kammoun, H. M.; Hamdaoui, B.; and Guizani, M. 2024. ChatGPT’s security risks and benefits: offensive and defensive use-cases, mitigation measures, and future implications. *IEEE Access*.
- Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3): 354–368.
- Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022a. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 427–439.
- Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022b. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 427–439.
- Dai, S.; Xu, C.; Xu, S.; Pang, L.; Dong, Z.; and Xu, J. 2024. Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6437–6447.

- Derczynski, L.; Kirk, H. R.; Balachandran, V.; Kumar, S.; Tsvetkov, Y.; Leiser, M. R.; and Mohammad, S. 2023. Assessing language model deployment with risk cards. *arXiv preprint arXiv:2303.18190*.
- Fagerland, M. W.; Lydersen, S.; and Laake, P. 2015. Recommended Confidence Intervals for Two Independent Binomial Proportions. *Statistical Methods in Medical Research*, 24(2): 224–254.
- Fang, H.; Miao, H.; Shukla, K.; Nanas, D.; Xu, C.; Greer, C.; Polyzotis, N.; Doshi, T.; Deng, T.; Mitchell, M.; et al. 2020. Introducing the model card toolkit for easier model transparency reporting. *Google AI Blog*.
- Girimella, K.; and Eckles, D. 2017. Image based misinformation on WhatsApp. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)*.
- Golda, A.; Mekonen, K.; Pandey, A.; Singh, A.; Hassija, V.; Chamola, V.; and Sikdar, B. 2024. Privacy and Security Concerns in Generative AI: A Comprehensive Survey. *IEEE Access*, 12: 53485–53499.
- Golpayegani, D.; Hupont, I.; Panigutti, C.; Pandit, H. J.; Schade, S.; O’Sullivan, D.; and Lewis, D. 2024. AI cards: towards an applied framework for machine-readable AI and risk documentation inspired by the EU AI Act. In *Annual Privacy Forum*, 48–72. Springer.
- Hameleers, M.; Powell, T. E.; Van Der Meer, T. G.; and Bos, L. 2020. A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political communication*, 37(2): 281–301.
- Herdel, V.; Šćepanović, S.; Bogucka, E.; and Quercia, D. 2024. ExploreGen: Large language models for envisioning the uses and risks of AI technologies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 584–596.
- Jaakkola, M. 2023. A handbook for journalism educators: Reporting on Artificial Intelligence. *United Nations Educational, Scientific and Cultural Organization*.
- Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3): e1452.
- Liang, W.; Rajani, N.; Yang, X.; Ozoani, E.; Wu, E.; Chen, Y.; Smith, D. S.; and Zou, J. 2024. Systematic Analysis of 32,111 AI Model Cards Characterizes Documentation Practice in AI. *Nature Machine Intelligence*, 6(7): 744–753.
- Liu, A.; Sheng, Q.; and Hu, X. 2024. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3001–3004.
- Liu, J.; Li, W.; Jin, Z.; and Diab, M. 2024a. Automatic Generation of Model and Data Cards: A Step Towards Responsible AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1975–1997.
- Liu, Y.; Huang, J.; Li, Y.; Wang, D.; and Xiao, B. 2025. Generative AI Model Privacy: A Survey. *Artificial Intelligence Review*, 58(33): 1–25.
- Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024b. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models’ Alignment. *arXiv:2308.05374*.
- Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36.
- McGregor, S. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15458–15463.
- McLean, S.; Read, G. J.; Thompson, J.; Baber, C.; Stanton, N. A.; and Salmon, P. M. 2023. The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(5): 649–663.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. Atlanta GA USA: ACM. ISBN 978-1-4503-6125-5.
- Mittelstadt, B.; Wachter, S.; and Russell, C. 2023. To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour*, 7(11): 1830–1832.
- Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2014–2037*. Dubrovnik, Croatia: Association for Computational Linguistics.
- NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).
- NIST. 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.
- Organisation for Economic Co-operation and Development (OECD). 2025. OECD AI Incidents Database. Online at <https://oecd.ai/en/incidents-methodology>. Accessed: August 2025.
- Ousidhoum, N.; Zhao, X.; Fang, T.; Song, Y.; and Yeung, D.-Y. 2021. Probing Toxic Content in Large Pre-Trained Language Models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4262–4274. Online: Association for Computational Linguistics.
- O’Connor, S.; and Liu, H. 2024. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & SOCIETY*, 39(4): 2045–2057.
- Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

- Perset, K.; and Aranda, L. 2024. Defining AI Incidents and Related Terms. OECD Artificial Intelligence Papers No. 16, OECD. Approved by the OECD Digital Policy Committee on 14 March 2024.
- Pownall, C. 2023. AI, Algorithmic and Automation Incident and Controversy Repository (AIAAIC). Online at <https://www.aiaaic.org/>. Accessed: August 2025.
- Raji, I. D.; and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.
- Rao, P. S. B.; Šćepanović, S.; Zhou, K.; Bogucka, E. P.; and Quercia, D. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.
- Rodrigues, R.; Resseguier, A.; and Santiago, N. 2023. When Artificial Intelligence Fails: The Emerging Role of Incident Databases. *Pub. Governance, Admin. & Fin. L. Rev.*, 8: 17.
- Salehzadeh Niksirat, K.; Goswami, L.; S. B. Rao, P.; Tyler, J.; Silacci, A.; Aliyu, S.; Aeblí, A.; Wacharamantham, C.; and Cherubini, M. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–23. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.
- Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.
- Scherer, R. 2018. PropCIs: Various Confidence Interval Methods for Proportions.
- Schmitt, M.; and Flechais, I. 2024. Digital deception: generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57(12): 324.
- Shaikh, S. J.; and Moran, R. E. 2024. Recognize the bias? News media partisanship shapes the coverage of facial recognition technology in the United States. *New Media & Society*, 26(5): 2829–2850.
- Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.
- Slattery, P.; Saeri, A. K.; Grundy, E. A.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. *AGI-Artificial General Intelligence-Robotics-Safety & Alignment*, 1(1).
- Sokol, A.; Moniz, N.; Daly, E.; Hind, M.; and Chawla, N. 2024. BenchmarkCards: Large Language Model and Risk Reporting. *arXiv preprint arXiv:2410.12974*.
- Stahl, B. C.; and Eke, D. 2024. The ethics of ChatGPT—Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74: 102700.
- Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11.
- Tahaei, M.; Constantinides, M.; Quercia, D.; Kennedy, S.; Muller, M.; Stumpf, S.; Liao, Q. V.; Baeza-Yates, R.; Aroyo, L.; Holbrook, J.; Luger, E.; Madaio, M.; Blumenfeld, I. G.; De-Arteaga, M.; Vitak, J.; and Olteanu, A. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394222.
- Turri, V.; and Dzombak, R. 2023. Why we need to know more: Exploring the state of AI incident documentation practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 576–583.
- Uuk, R.; Brouwer, A.; Schreier, T.; Dreksler, N.; Pulignano, V.; and Bommasani, R. 2024. Effective Mitigations for Systemic Risks from General-Purpose AI. *arXiv preprint arXiv:2412.02145*.
- Velázquez, J. D. M.; Šćepanović, S.; Gvirtz, A.; and Quercia, D. 2024. Decoding Real-World Artificial Intelligence Incidents. *Computer*, 57(11): 71–81.
- Wang, Z.; Zhu, H.; Liu, P.; and Sun, L. 2021. Social engineering in cybersecurity: a domain ontology and knowledge graph application examples. *Cybersecurity*, 4(1): 31.
- Wang, Z. J.; Kulkarni, C.; Wilcox, L.; Terry, M.; and Madaio, M. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–40.
- Washo, A. H. 2021. An interdisciplinary view of social engineering: A call to action for research. *Computers in Human Behavior Reports*, 4: 100126.
- Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
- Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.
- Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 214–229. New York,

NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Whittaker, M.; Alper, M.; Bennett, C. L.; Hendren, S.; Kazianas, L.; Mills, M.; Morris, M. R.; Rankin, J.; Rogers, E.; Salas, M.; et al. 2019. Disability, bias, and AI. *AI Now Institute*, 8.

Wirtz, B. W.; Weyrer, J. C.; and Sturm, B. J. 2020. The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration*, 43(9): 818–829.

Yampolskiy, R. V. 2016. Taxonomy of pathways to dangerous artificial intelligence. In *Workshops at the thirtieth AAAI conference on artificial intelligence*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403.

Yu, Y.; Zhuang, Y.; Zhang, J.; Meng, Y.; Ratner, A. J.; Krishna, R.; Shen, J.; and Zhang, C. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.

Appendix

(A) AI Model Risks Catalog

Table 2: Categories of risks in the AI Model Risk Catalog (2024) tied to specific modalities. *Left:* The bars for each category show the percentage and number of risks linked to different types of input or output modalities such as images or text. *Right:* Two most frequently mentioned risks for each category. For the top risk in each category, we also show an example AI model associated with that risk. Most models use text, so most risks are linked to text too. But risks from malicious uses are showing up in models with multi-modal inputs. Risks about system safety also affect models with image and audio inputs.

Categories of risks per input/output modality type	Multimodal	Text	Audio	Image	Video	Most frequent consolidated risks	#
AI system safety, failures & limitations							
Output		10539			1593	Generates responses that are irrelevant or incorrect when interpreting complex, nuanced or ambiguous queries (Llama3-Ko-Carrot-8B-it)	21
Input		11146			62264	Contains bugs, inefficiencies and unexpected behaviour in generated code	19
Discrimination & toxicity							
Output		11833			2546	Perpetuates biases present in the training data (llama3-8b-italIA-unslot)	66
Input		13243			546	Inherits biases inherent in the training data	8
Overreliance & human agency issues							
Output		189				Increases risk in decision-making impacting individuals or society (AI-Buddy)	2
Input		181			65	Should not be considered a substitute for professional mental health support or counseling	2
Malicious actors & misuse							
Output		918			377	Engages in illegal or unethical activities if misused (TinyLlama-1.1bee)	5
Input		1050			185	Impersonates individuals or organizations without consent	4
Misinformation							
Output		2990			392	Generates factually incorrect information (mpt-7b)	13
Input		3315				Produces incorrect information as if it were factual	10
Privacy & security							
Output		575			339	Contains bugs and security vulnerabilities in generated code (starchat2-15b-v0.1)	6
Input		731			122	Violates privacy by exposing personally identifiable information	4
Socioeconomic & Environmental Harms							
Output		145			9	Requires significant computational resources and time (T0_3B)	3
Input		123			23	Requires significant resources for finetuning and inference	2

0% 25% 50% 75% 100%

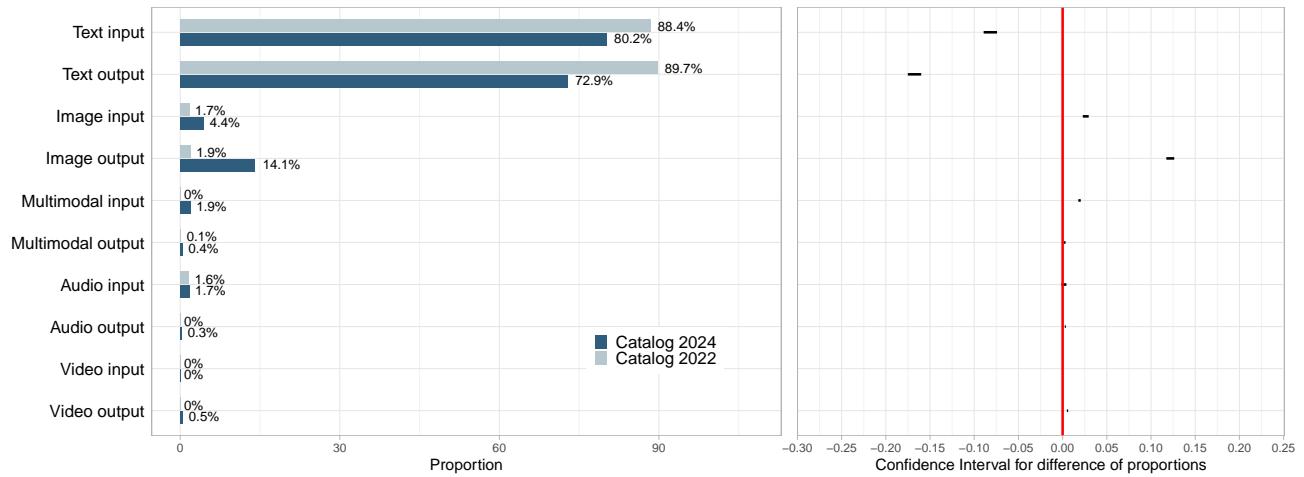


Figure 4: Comparing the AI Model Risk Catalogs from 2022 and 2024 by model input and output types. *Left:* The percentages of risks in the catalog that belong to each category. *Right:* 95% confidence intervals for the pairwise differences between each pairs of percentages. A greater distance of the interval from the vertical red line indicates a larger difference, while a narrower interval reflects higher certainty. Both text input and output have declined since 2022, but still make up 70–80% of modalities. This reflects a more than threefold rise in image inputs and outputs. Multimodal input has also grown quickly, from almost zero to 2%. These results show a clear trend toward more non-textual modalities (Yin et al. 2024).

(B) Comparing risks across sources

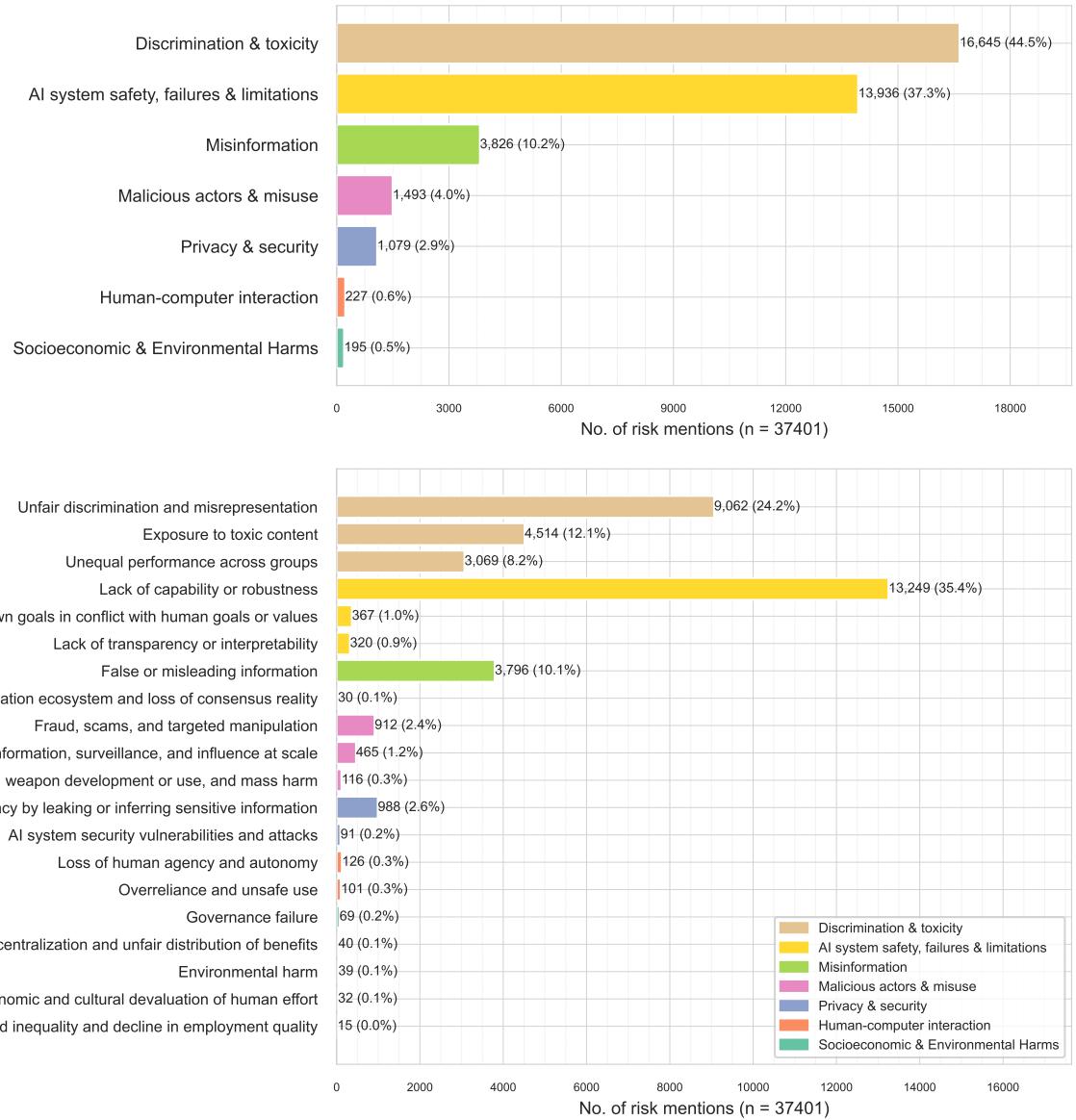


Figure 5: Proportions of risks in the *AI Model Risk Catalog* falling into different *MIT taxonomy* (Slattery et al. 2024) categories (top) and subcategories (bottom).

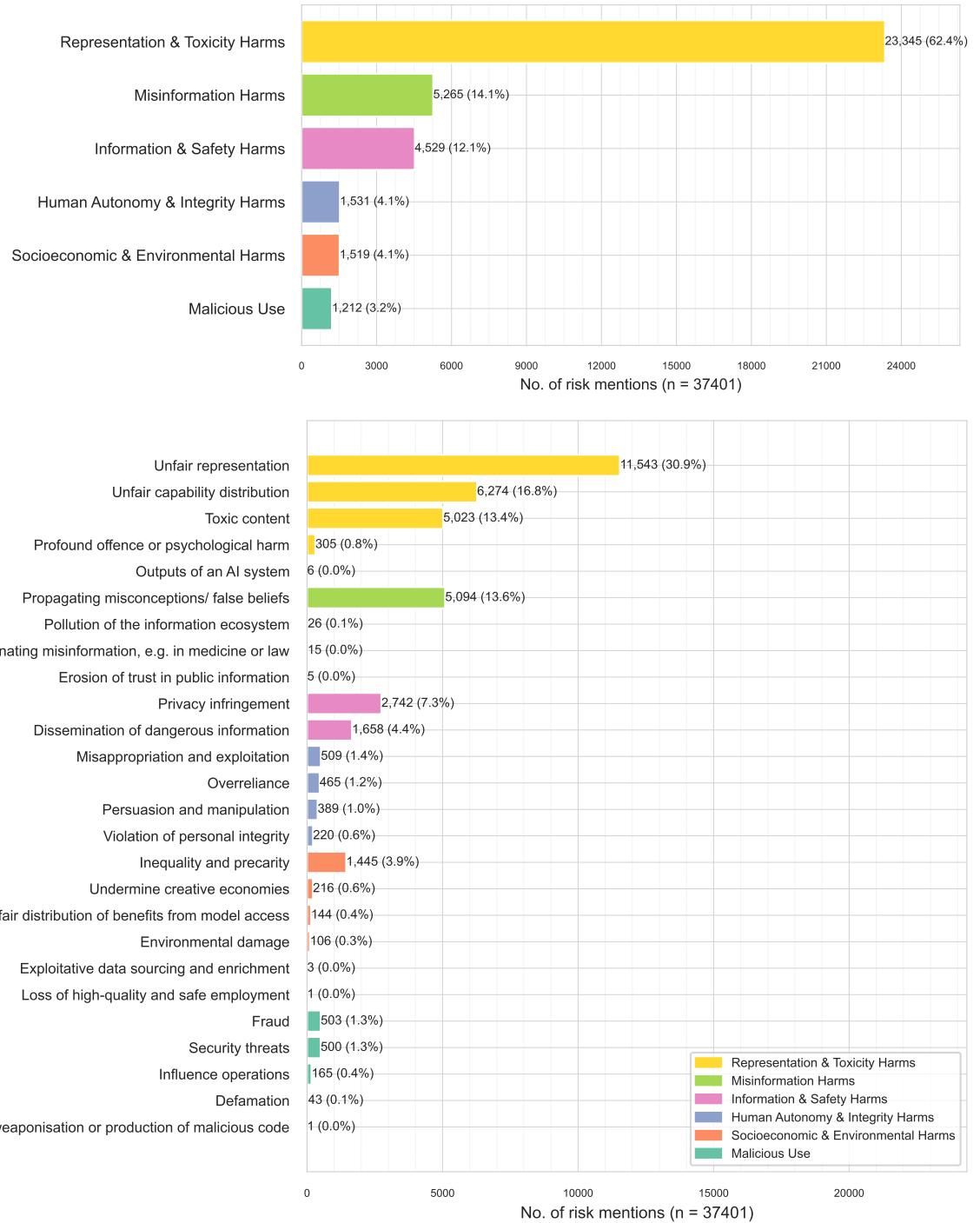


Figure 6: Proportions of risks in the *AI Model Risk Catalog* falling into different *DeepMind taxonomy* (Weidinger et al. 2022) categories (top) and subcategories (bottom).

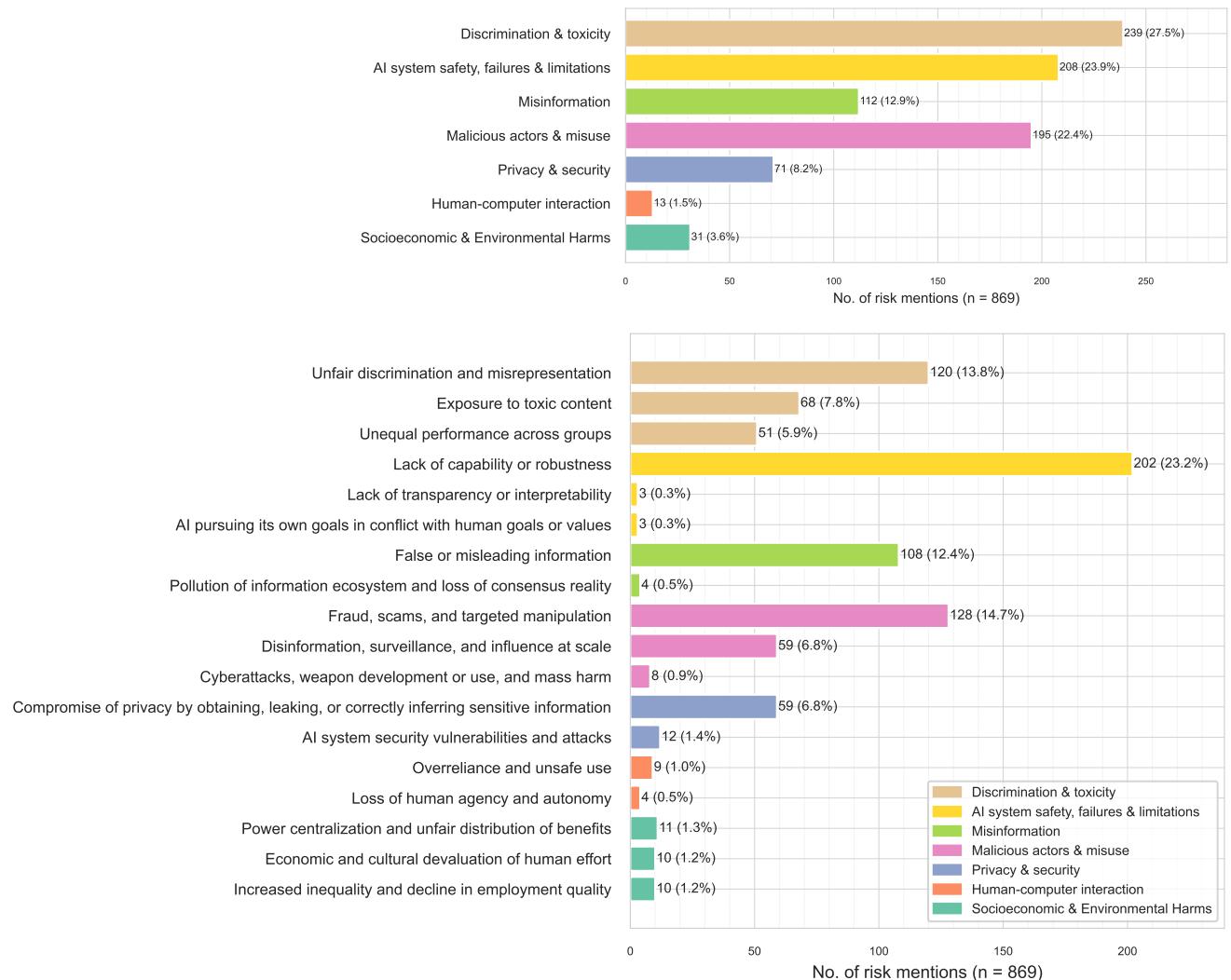


Figure 7: Proportions of risks in the *AI Incident Database* falling into different *MIT taxonomy* (Slattery et al. 2024) categories (top) and subcategories (bottom).



Figure 8: Proportions of risks in the *MIT Risk Repository* falling into different *MIT taxonomy* (Slattery et al. 2024) categories (top) and subcategories (bottom).

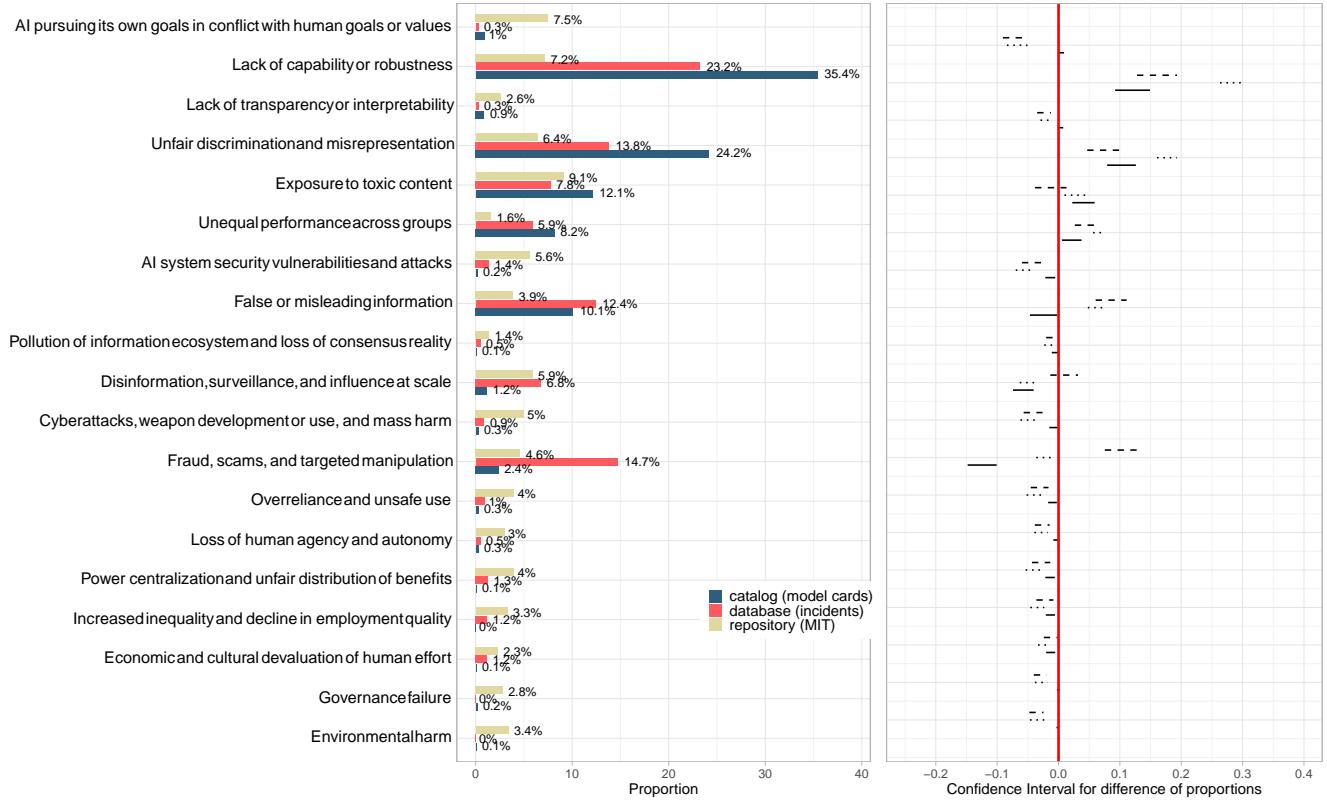


Figure 9: Comparison of three risk sources per MIT taxonomy *subcategory*. We compare our catalog (model cards), repository (MIT), and the database (AI Incidents). *Left:* The percentages of risks in the catalog / database / repository that belong to each subcategory. *Right:* 95% confidence intervals (CI) for the pairwise differences between each pairs of percentages. A greater distance of the interval from the vertical red line indicates a larger difference, while a narrower interval reflects higher certainty.

Prompts

Risk Extraction Prompt

1 System:

2 Consider the following definitions: 1) An AI incident is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following harms: (a) injury or harm to the health of a person or groups of people; (b) disruption of the management and operation of critical infrastructure; (c) violations to human rights or a breach of obligations under the applicable law intended to protect fundamental, labour and intellectual property rights; (d) harm to property, communities or the environment.¹ The harm can be physical, psychological, reputational, economic/financial (including harm to property), environmental, public interest (e.g., protection of critical infrastructure and democratic institutions), human rights and fundamental rights. 2) An AI risk is expressed as likelihood that harm or damage will occur. Risk is a function of both the probability of an event occurring and the severity of the consequences that would result. Risk is usually expressed in terms of risk sources, potential events, their consequences and their likelihood. \\

3 User: You are provided in input with cardID and a section written by an AI developer from Risks, Bias and Limitations section of an AI model card:

4

5 cardID: "{}", section: "{}".

6

7 Tasks:

8 1) Identify Relevance: Determine for each section whether it discusses the AI risks and potential solutions to mitigate them. Provide a response as Yes, No, or Unclear. If the answer is Yes, continue with the next steps; else, you will output empty "Risks and Mitigations" in the output JSON.

9 %

10 2) Analyze: Identify all the subtexts in the section that can be categorized into - "RISKS", "MITIGATIONS", or "MIX OF BOTH". Classify each such subtext into one of the provided three types. This subtext must match the exact sentences in the input section and is used as a reference.

11 %

12 3) Split: For each reference, identify unique risks and mitigations. Classify each into one of the two types "RISKS" or "MITIGATIONS". It is very important that each risk or mitigation identified be formatted like this: Verb + Object + [Explanation], and describes one unique risk/mitigation, is concise, consisting of one clear, to-the-point sentence, with up to maximum of 20 words. Specifically, start a risk with an action verb in active present tense (e.g., undermines, discriminates, infringes, reduces, increases etc., but NOT potentially) followed by the object and the reason in case it is not obvious and requires an explanation. Start a mitigation with verb in present tense, base form (e.g., perform, inform, investigate). Also, we want these descriptions to be read by broad public who does not have a deep knowledge about AI technology. Do not change the meaning of the concrete risk. For example, if it says the model does not work in a particular setting then you say it underperforms and not discriminates.

13

14 Good and Bad Examples: This would NOT be a good output: "Potentially undermines the right to privacy if the facial recognition data is not properly secured." while this is good: "Undermines the right to privacy if the facial recognition data is not properly secured.". This would NOT be a good output: "Affects all fine tuned versions of the model." as it does not specify what the risk is; instead "Transfers all the <specific risks from reference like bias, unrepresentative data> to the inherited models" is better. "Evaluations surface potential risks in use cases." is a bad output, instead specify the potential risks from the reference.

15 Other examples of well-formatted risks:

16 * Underperforms on non-English languages

17 * Discriminates against certain players, such as women or those from certain ethnic backgrounds.

18 * Undermines the safety and security when model is deployed without thorough in-domain testing.

19 Examples of well-formatted mitigations:

20 * Perform thorough in-domain testing

21 * Update model card for models using pretrained models

22 * Inform indirect users when the content they're working with is created by the LLM.

Risk Classification Prompt – MIT

```
1 MESSAGES = [ { 'role': 'system', 'content': """As a distinguished expert in artificial intelligence technology, you embody the
2   forefront of Responsible Artificial Intelligence (RAI). Your expertise is not just technical; it's deeply rooted in a
3   conscientious approach to AI's ethical, social, and environmental implications. With a wealth of experience, you navigate
4   the intricate balance between harnessing AI's potential for positive impact and mitigating its risks. Consider the
5   following definitions:
6
7   1) An AI incident is an event, circumstance or series of events where the development, use or malfunction of one or more AI
8     systems directly or indirectly leads to any of the following harms: (a) injury or harm to the health of a person or
9     groups of people; (b) disruption of the management and operation of critical infrastructure; (c) violations to human
10    rights or a breach of obligations under the applicable law intended to protect fundamental, labour and intellectual
11    property rights; (d) harm to property, communities or the environment.' The harm can be physical, psychological,
12    reputational, economic/financial (including harm to property), environmental, public interest (e.g., protection of
13    critical infrastructure and democratic institutions), human rights and fundamental rights.
14
15   2) An AI risk is expressed as likelihood that harm or damage will occur. Risk is a function of both the probability of an
16    event occurring and the severity of the consequences that would result. Risk is usually expressed in terms of risk
17    sources, potential events, their consequences and their likelihood."""
18
19   3) Tasks:
20     1. Classify the risk along this axis:
21
22     * Axis: 1. Discrimination \& toxicity, 2. Privacy \& security, 3. Misinformation, 4. Malicious actors \& misuse, 5. Human-
23       computer interaction, 6. Socioeconomic \& Environmental Harms 7. AI system safety, failures \& limitations
24
25   4) For the axis, they are distinguished by the areas of risks causing potential harm:
26
27   5) <Placeholder for definitions borrowed from the MIT AI risk repository>
28
29   6) Important Note:
30
31     1) Each incident should be classified under one of the seven risk areas without overlap. Under each risk area, the incident needs to
32       be further classified under their respective sub-categories. Each incident under a risk area should be classified ONLY
33       into sub-categories under that risk area.
34
35   7) Output Format: Ensure your output strictly follows this JSON structure.
36
37   8) {
38     "id": "<id>",
39     "risk": "<risk>",
40     "axis2_risk_area": [
41       {
42         "risk_area": one of ["Discrimination \& toxicity", "Privacy \& security", "Misinformation", "Malicious actors & misuse
43           ", "Human-computer interaction", "Socioeconomic \& Environmental Harms", "AI system safety, failures \&
44           limitations"],
45         "sub_category": ["one of the sub-categories of the above-classified risk area"]
46       },
47     ],
48   }
49
50   9) Important Notes: Do not report your reasoning steps or any preamble like 'Here is the output' or 'JSON', ONLY the JSON result. In
51       scenarios where no sentences are mentioned, provide an empty JSON array.
52
53   10) *** Double-check your output to ensure it contains only the requested JSON and nothing else. *** """
54 }
```

Risk Classification Prompt – DeepMind

1 System: As a distinguished expert in artificial intelligence technology, you embody the forefront of Responsible Artificial Intelligence (RAI). Your expertise is not just technical; it's deeply rooted in a conscientious approach to AI's ethical, social, and environmental implications. With a wealth of experience, you navigate the intricate balance between harnessing AI's potential for positive impact and mitigating its risks. Consider the following definitions: 1) An AI incident is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following harms: (a) injury or harm to the health of a person or groups of people; (b) disruption of the management and operation of critical infrastructure; (c) violations to human rights or a breach of obligations under the applicable law intended to protect fundamental, labour and intellectual property rights; (d) harm to property, communities or the environment.' The harm can be physical, psychological, reputational, economic/financial (including harm to property), environmental, public interest (e.g., protection of critical infrastructure and democratic institutions), human rights and fundamental rights. 2) An AI risk is expressed as likelihood that harm or damage will occur. Risk is a function of both the probability of an event occurring and the severity of the consequences that would result. Risk is usually expressed in terms of risk sources, potential events, their consequences and their likelihood.

2 \\

3 User: You are provided in input with a sentence describing AI risk and its ID - id: "{}", risk: "{}".

4

5 Tasks:

6 Group these risks along two axes:

7

8 * Axis 1: 1. Capability; 2. Human Interaction; And 3. Systemic.

9

10 * Axis 2: 1. Representation \& Toxicity Harms, 2. Misinformation Harms, 3. Information \& Safety Harms, 4. Malicious Use, 5. Human Autonomy \& Integrity Harms, 6. Socioeconomic \& Environmental Harms

11

12 For the first axis, the three evaluation layers for risks are distinguished by the target of analysis.

13

14 (1) Capability: targets AI systems, their technical components, and the processes by which these systems and components are created, including:

15 (a) Outputs of an AI system, for example, model performance, efficiency metrics such as energy use, the extent to which an AI model reproduces harmful stereotypes, factual errors, or displays advanced capabilities that present safety hazards,

16 (b) The data on which a model is trained, for example, the diversity and representativeness of the data, the presence of sensitive data, the learned associations of a trained AI system,

17 (c) Filters and techniques for reducing system harms, such as filters for toxic language.

18

19 (2) Human Interaction: targets the experience of people interacting with AI systems and their effects on these people, including:

20 (a) The system's usability, for example, whether the AI system performs its intended function at the point of use, how experiences differ between user groups, and how easy it is to use a model for malicious ends,

21 (b) Potential externalities, for example, whether human-AI interaction leads to unintended effects on the person interacting with or exposed to AI outputs, such as overreliance on AI systems, overtrust, and cognitive biases,

22 (c) Potential harms, including harms to data annotators and harms arising from different system modalities (e.g., video, image , text),

23 (d) The overall quality of outcomes in human-AI assisted tasks compared to human-human assisted tasks.

24

25 (3) Systemic Impact: targets the impact of an AI system on the broader systems in which it is embedded, such as society, the economy, and the natural environment, including:

26 (a) Systems of various domains, sizes, industries, or goods,

27 (b) Adoption and perception of AI across different systems,

28 (c) Distribution of benefits and risks from the AI,

29 (d) Environmental impacts of the AI on the systems, e.g., biodiversity and resilience of local ecosystems.

30

31 For the second axis, they are distinguished by the areas of risks causing potential harm.

32

33 <Placeholder for definitions borrowed from the DeepMind taxonomy>

34

35 Important Note: 1) If a risk is initially categorised under 'Capability' and 'Human Interaction' or any other overlapping categories from Axis 1, select the most appropriate single category. Exclude it from any additional categories by maintaining a distinct risk set.

36 2) For axis 2, each risk should be classified under one of the six risk areas without overlap. Under each risk area, the risk needs to be further classified under their respective sub-categories. Each risk under a risk area should be classified ONLY into sub-categories under that risk area.

37 3) Classify the subcategories under the respective axis only. DO NOT mix them across the axes.